**一、数据清洗及导入**  
 本次数据分析采用了UCI机器学习数据库中的apartment + for + rent + classified数据集作为分析对象。

对数据集进行初步了解可知，其中包含的属性有：ID; 类别; 标题; 内容; 设施; 浴室数; 卧室数; 租金货币类型; 有无额外缴费; 有无照片介绍; 是否允许饲养宠物; 价格; 网站显示的租金; 缴费类型（周or月）;房屋面积; 地址; 城市名称; 州; 纬度; 经度; 源; 时间。

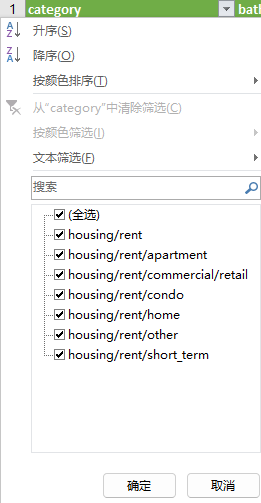
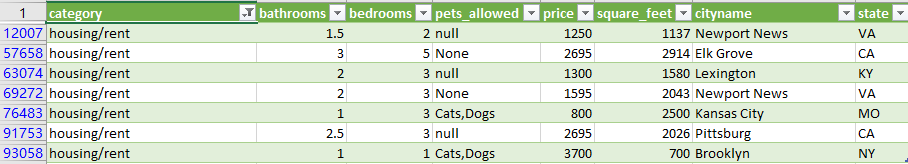
首先明确本次建模的目标：利用数据集内提供的房屋信息，建立一个租金回归预测模型，以及一个类别分类模型。

接下来对数据的属性做出取舍，去掉对建模无益的数据以及过于繁杂的数据种类，简化一部分建模的流程。首先针对ID，上传的标题，内容，经纬度，地址，网站来源，上传时间这类用于辨识数据条目的“独特”属性种类，予以去除。有一些属性值，在整个数据集内仅有一类完全相同的值（如租金货币仅包含美金），包括租金货币类型，有无额外缴费（数据集内均为“无”），也予以去除。价格和网站显示的租金内容重复，将属性“网站显示的租金”去除。另有“有无照片介绍”，“缴费类型”，“设施”等属性值出现过于单一，缺乏分类意义或缺失值过多的，也不纳入建模的考虑过程。还有“城市”这种取值过多的，会影响回归模型的建立，也需要去除。最终，仅保留“类别”“浴室数”，“卧室数”，“是否允许饲养宠物”，“价格”，“房屋面积”作为纳入建模的几大参考属性。

然后针对离群值和缺失值进行处理，这里依次对每个属性进行展开叙述。

**1.Categories**

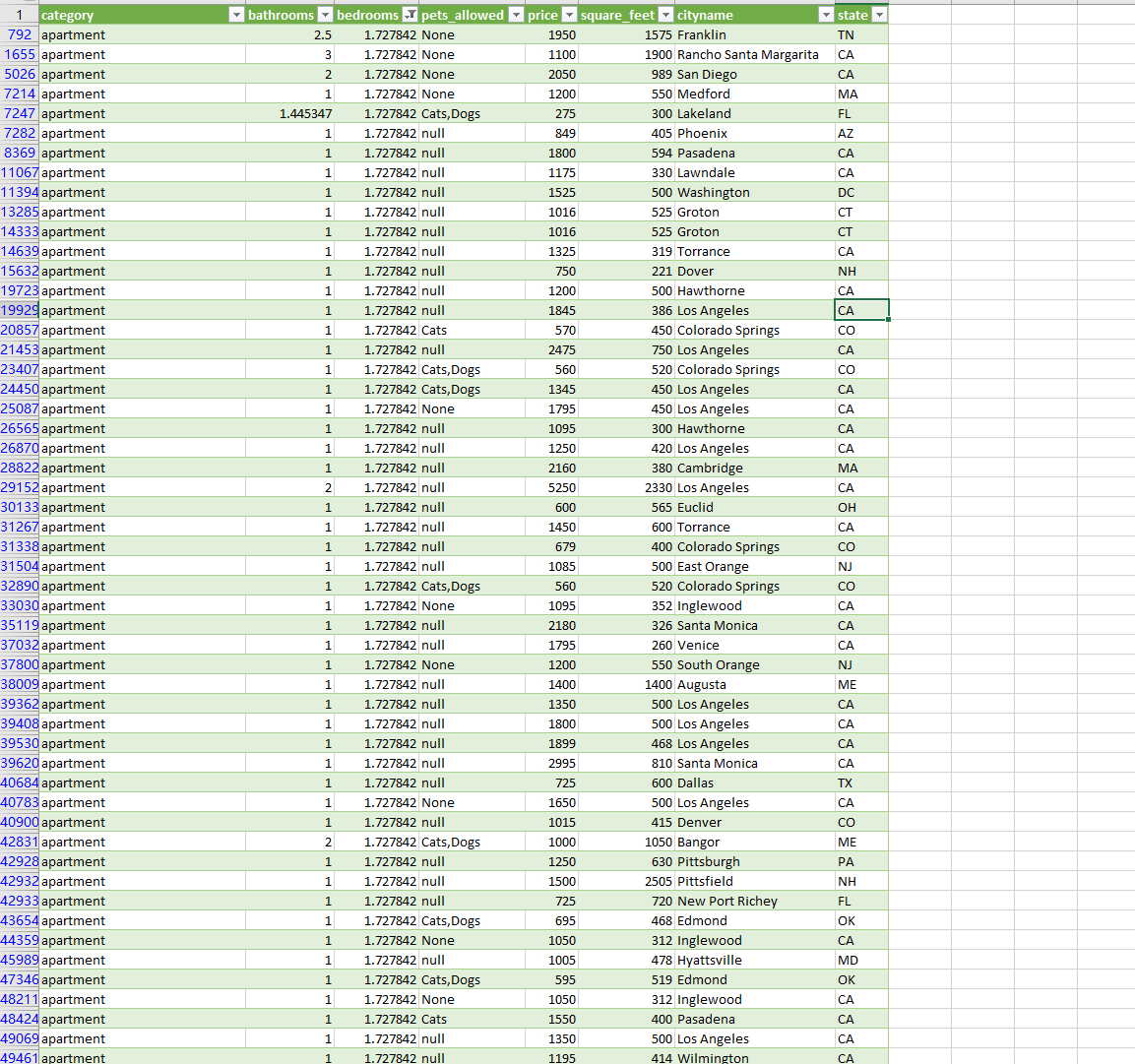
首先是类别属性。可见图中分出了7个种类。

首先看到rent分类，这些是没有明确标定分类的出租房屋，仅有7条数据，比例非常少，可以按缺失值处理，直接移除。

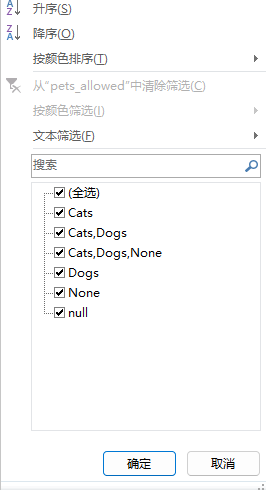
Apartment和Commercial retail分别代表公寓出租和商业网点（门头房）出租，为可视化结果输出方便，可以将分类重命名，只保留它们的种类而非原数据的路径。另有Condo类与Apartment种类相近，都是公寓房，因此将Condo类数据合并入Apartment类。

最后，有others类、home类、short\_term类数据过少，可能影响将来的分类预测算法作为离群值去除。

**2.Bathrooms & Bedrooms**

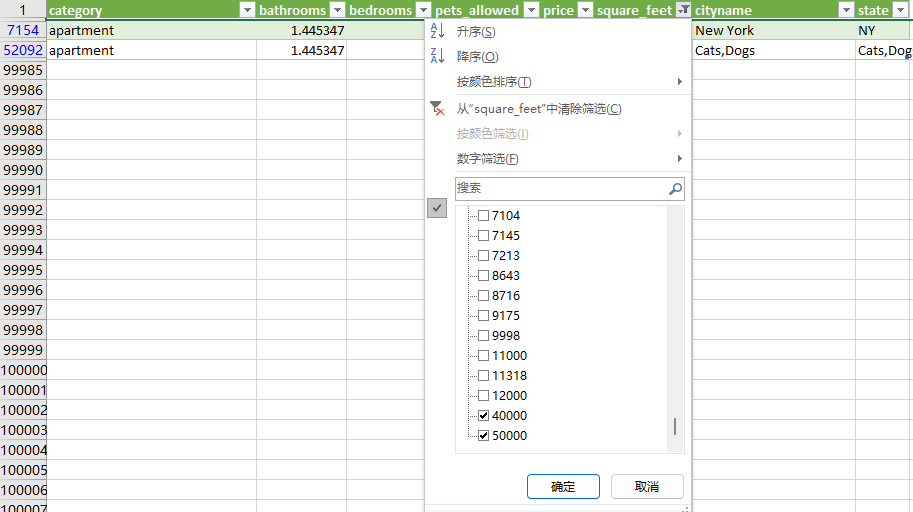
这一种类包含出租房屋内卫生间数量和卧室的数量，属于离散化的数据。这里在EXCEL中处理数据时发现有一部分缺失值。挑选出整行空白的缺失数据，去除；并使用均值填补剩余的缺失处。

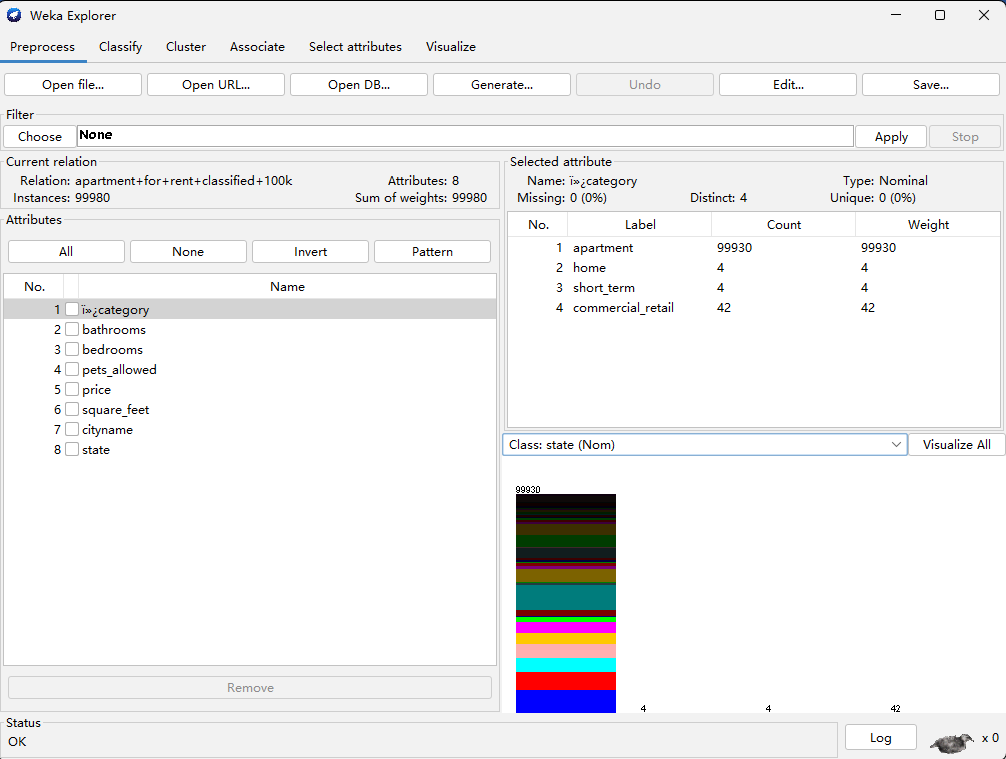
**3.Pets\_allowed**

这个属性包含是否允许饲养宠物。值包括猫；狗；猫和狗；猫、狗和无；无；缺失值。其中无代表不允许，还有一条数据是猫、狗和无，属于离群值，去除。

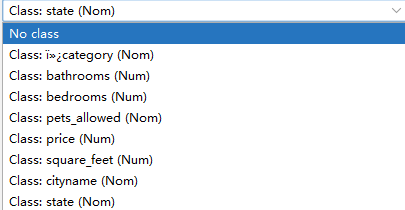
针对缺失值，因为其占比已有55%过半（55565 out of 99984），又考虑到剩下类别中“猫和狗”占绝对多数（37276 out of 99984）这里采取众数填充法，将其补充为cats and dogs.

**4. Price & Square\_feet**

这类别里是房屋的租金价格和房屋面积。挑选出离群值和缺失值去除即可。

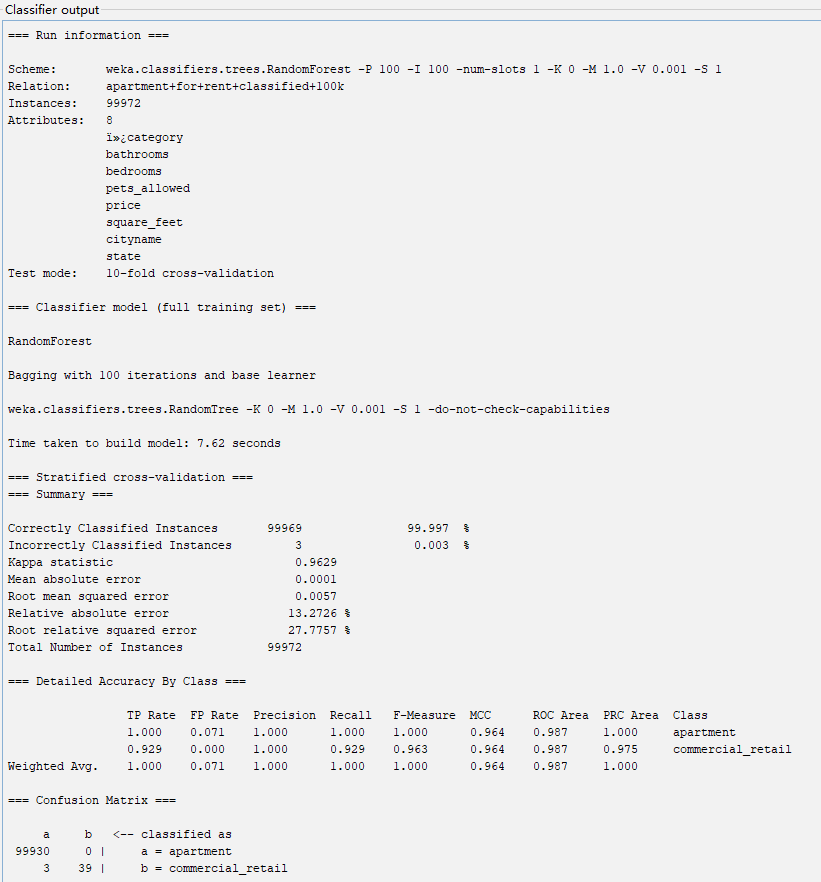
使用WEKA导入已经清洗完成的csv数据集文件。

首先对各属性的**数据类型Attribute Type**进行检查。

* **Category:** Nominal
* **Bathrooms:** Numeric
* **Bedrooms:** Numeric
* **Pets Allowed:** Nominal
* **Price:** Numeric
* **Size:** Numeric
* **City:** Nominal
* **State:** Nominal

其中类别、养宠情况、城市和州都属于标称属性；而剩下的都是数字属性。

**二、分类建模** 如果要对**Categories**进行分类建模预测，考虑到数据集中出现了严重的不平衡现象，可能会导致建模对少数属性的预测效果极差，所以不考虑使用朴素贝叶斯算法，转而考虑使用对不平衡数据更鲁棒的算法，这里采用随机森林+10折交叉验证建模。

Output:  
针对此结果进行了评估：

**1. 总体性能 (Summary)**

* **Correctly Classified Instances: 99.997%** - 准确率极高，接近完美。
* **Incorrectly Classified Instances: 0.003%** - 错误率极低。
* **Kappa statistic: 0.9629** - Kappa 统计量非常高，接近于 1，表明模型的一致性非常好，远优于随机猜测。
* **Mean absolute error (MAE): 0.0001** - 平均绝对误差非常小。
* **Root mean squared error (RMSE): 0.0057** - 均方根误差也很小。
* **Relative absolute error: 13.2726%** - 相对绝对误差很低。
* **Root relative squared error: 27.7757%** - 均方根相对误差相对朴素贝叶斯来说有所降低。

**2. 分类准确率细节 (Detailed Accuracy By Class)**

| **Class** | **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **ROC Area** | **PRC Area** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| apartment | 1.000 | 0.071 | 1.000 | 1.000 | 1.000 | 0.964 | 0.987 | 1.000 |
| commercial\_retail | 0.929 | 0.000 | 1.000 | 0.929 | 0.963 | 0.964 | 0.987 | 0.975 |
| **Weighted Avg.** | 1.000 | 0.071 | 1.000 | 1.000 | 1.000 | 0.964 | 0.987 | 1.000 |

* **apartment:** 表现完美。
* **commercial\_retail:** 召回率 (TP Rate) 为 0.929，精确率 (Precision) 为 1.000，F1 值也达到了 0.963。这意味着模型几乎没有将 "apartment" 误判为 "commercial\_retail"，并且正确识别了绝大多数 "commercial\_retail" 实例。

**3. 混淆矩阵 (Confusion Matrix)**

a b <-- classified as

99930 0 | a = apartment

3 39 | b = commercial\_retail

* 所有 "apartment" 类别的实例都被正确分类。
* 只有 3 个 "commercial\_retail" 类别的实例被错误地分类为 "apartment"。

**三、回归预测**

本研究旨在利用机器学习方法，对公寓出租信息数据集进行分析，并建立一个预测公寓租金（price）的模型。我们选择随机森林（Random Forest）作为主要的回归算法。随机森林是一种集成学习方法，理论上能够处理大规模数据、非线性关系和特征交互，并且对异常值具有较好的鲁棒性。然而，本次实验的结果显示模型性能不佳，需要进行详细分析和改进。

**1. 数据集**

本研究使用的数据集为公寓出租信息数据集，包含以下特征：

* bathrooms (浴室数量)
* bedrooms (卧室数量)
* pets\_allowed (是否允许宠物)
* price (公寓租金，**目标变量**)
* square\_feet (公寓面积)

数据集包含99971条实例（在本次运行中，有一条数据可能因为缺失值或其他原因被移除）。

**2. 实验设置**

* **工具：** WEKA (Version 3.8.6)
* **算法：** 随机森林 (RandomForest)
* **参数：**
  + -P 100: 使用100%的数据进行自助采样（bootstrap sampling）。
  + -I 100: 构建100棵决策树。
  + -num-slots 1: 使用1个线程。
  + -K 0: 每个节点分裂时考虑的特征数量（0表示使用所有特征数量的平方根）。
  + -M 1.0: 叶节点所需的最小实例数。
  + -V 0.001: 剪枝方差的最小比例。
  + -S 1: 随机数种子。
* **评估方式：** 10折交叉验证 (10-fold cross-validation)

**3. 实验结果**

WEKA随机森林回归模型的运行结果如下：

=== Run information ===

Scheme: weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Relation: apartment+for+rent+classified+100k-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R6-weka.filters.unsupervised.attribute.Remove-R6

Instances: 99971

Attributes: 5

bathrooms

bedrooms

pets\_allowed

price

square\_feet

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 15.97 seconds

=== Cross-validation ===

=== Summary ===

Correlation coefficient 0.4609

Mean absolute error 492.6937

Root mean squared error 816.2084

Relative absolute error 90.2769 %

Root relative squared error 90.3675 %

Total Number of Instances 99971

**4. 结果分析**

* **模型构建时间：** 15.97秒。
* **评估指标：**

相关系数 (Correlation coefficient): 0.4609

相关系数衡量了预测值与实际值之间的线性相关程度。0.4609 表明模型预测值与实际租金价格之间存在中等程度的正相关。这意味着模型的预测趋势与实际情况基本一致，但预测精度有限。

平均绝对误差 (MAE): 492.6937

MAE 表示预测值与实际值之间的平均绝对差异。492.6937 的 MAE 表明，模型的平均预测误差约为 492.7 美元。

均方根误差 (RMSE): 816.2084

RMSE 是预测误差的标准差，对较大的误差更敏感。816.2084 的 RMSE 表明模型预测存在较大的波动性，部分预测值可能与实际值偏差较大。

相对绝对误差 (RAE): 90.2769%

RAE 将模型的 MAE 与一个简单基准模型（通常是使用平均值进行预测）的 MAE 进行比较。90.2769% 的 RAE 表明模型的预测误差接近于基准模型的误差，说明模型的预测能力较弱。

相对均方根误差 (RRSE): 90.3675%

RRSE 与 RAE 类似，但使用的是 RMSE。90.3675% 的 RRSE 同样表明模型的预测能力较弱。

均方误差 (MSE): 666196.14

MSE 表示预测误差的平方的平均值。MSE 值较大，表明模型预测存在较大的误差。

总平方和(SSE): 66,599,995,753.54

SSE值较大，表明模型预测存在较大的误差。

**5. 可视化分析**

我们使用WEKA的“Visualize”面板对模型的预测结果进行了可视化分析。

* **X轴：predicted price（预测价格），Y轴：price（真实价格）**
  + **理想情况：** 数据点应该紧密地分布在对角线（y=x）附近，表示预测值与真实值非常接近。
  + **实际情况：** 散点图显示数据点分布大致关于y=x对称，虽然从图上看对称性不错但是误差并不算小。
  + **观察：** 可以看出模型倾向于低估高价公寓的价格，高估低价公寓的价格。

**可视化分析**

1. **数据分布:**
   * 大部分数据点集中在图的左下角，表明大多数公寓的实际租金和预测租金都相对较低。
   * 存在一些离群点（outliers），位于图的右上角，这些公寓的实际租金和预测租金都非常高。
   * 数据点呈现出一定的正相关趋势，即预测价格越高，实际价格也倾向于越高，但相关性并不强，点比较分散。
2. **模型表现:**
   * 理想情况下，如果模型的预测完全准确，所有的数据点都应该落在一条从左下角到右上角的对角线上（y=x）。
   * 从图中可以看出，数据点偏离这条对角线较远，表明模型的预测存在较大误差。
   * 尤其是在高价区域（图的右上角），模型的预测似乎更加不准确，存在明显的低估现象。
   * 许多预测值聚集在较低的数值上，而实际值具有更广的范围，表明模型的预测范围受限。
3. **误差模式:**
   * 观察垂直方向的偏差：对于给定的预测价格（x值），实际价格（y值）的分布范围很广，表明模型对同一预测价格的公寓，其真实价格的预测不确定性很高。
   * 观察水平方向的偏差：对于给定的实际价格（y值），预测价格（x值）也有一定范围，表明模型对具有相同真实价格的公寓，预测价格也不一致。

**综合结论**

基于图片的可视化分析，可以得出以下结论：

* 随机森林模型在预测公寓租金价格方面表现**不佳**，与之前实验报告中的数值结果一致。
* 模型预测值与实际值之间存在偏差，预测精度有限。
* 模型在预测高价公寓时，存在明显的低估现象。
* 模型的预测范围受限，不能很好地反映实际租金价格的分布。
* 可视化结果进一步证实了需要改进模型的必要性。

**四、聚类分析**

本部分旨在利用聚类分析技术，对公寓出租信息数据集进行更深入的探索，以识别潜在的公寓类型或租户群体。考虑到数据集的规模和潜在的内存限制，我们选择EM（Expectation-Maximization）算法作为主要的聚类方法。EM算法是一种基于概率模型的聚类算法，能够有效地处理大规模数据，并对数据分布进行建模。

**1. 实验设置**

* **工具：** WEKA (Version 3.8.6)
* **算法：** EM
* **数据集：** 经过预处理的公寓出租信息数据集
* **属性：**
  + bathrooms (浴室数量)
  + bedrooms (卧室数量)
  + square\_feet (面积)
* **EM算法参数：**
  + -I 100: 最大迭代次数（100次）
  + -N -1: 自动确定簇的数量（通过交叉验证）
  + -X 10: 交叉验证折数（10折）
  + -max -1: 不限制最大实例数
  + -ll-cv 1.0E-6: 交叉验证对数似然阈值
  + -ll-iter 1.0E-6: 迭代对数似然阈值
  + -M 1.0E-6: 最小标准差
  + -K 10: 内部KMeans算法的簇数量（用于初始化）
  + -num-slots 1: 使用的线程数
  + -S 100: 随机数种子

**2. 实验结果**

WEKA EM算法的运行结果如下：

=== Run information ===

Scheme: weka.clusterers.EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100

Relation: apartment+for+rent+classified+100k-weka.filters.unsupervised.attribute.Remove-R1,5,7-8-weka.filters.unsupervised.attribute.Remove-R3

Instances: 99972

Attributes: 3

bathrooms

bedrooms

square\_feet

Test mode: evaluate on training data

=== Clustering model (full training set) ===

EM

==

Number of clusters selected by cross validation: 3

Number of iterations performed: 1

Cluster

Attribute 0 1 2

(0.55) (0.13) (0.32)

============================================

bathrooms

mean 1 2.0094 1.9795

std. dev. 0 0.6083 0.1493

bedrooms

mean 1.26 2.9975 2

std. dev. 0.4785 0.7372 0.0027

square\_feet

mean 750.0475 1462.7267 1096.4291

std. dev. 165.9468 595.499 154.3346

Time taken to build model (full training data) : 37.36 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 56268 ( 56%)

1 12118 ( 12%)

2 31586 ( 32%)

Log likelihood: 0.15434

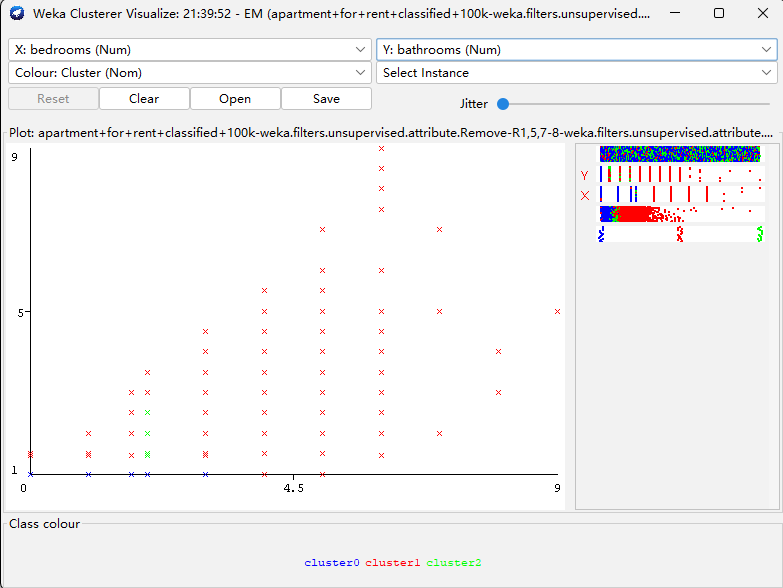
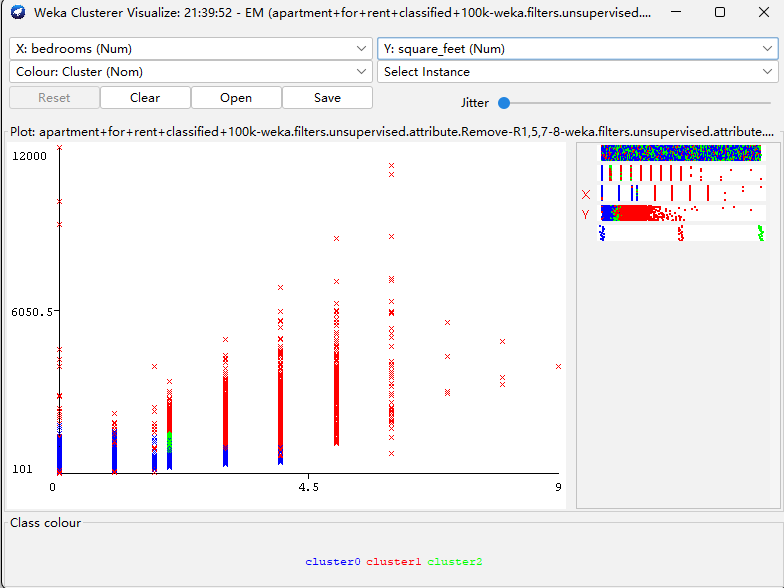
**3. 结果分析**

EM算法通过交叉验证确定了3个簇，分别代表了不同类型和规模的公寓：

* **簇0 (56%)：小户型公寓**
  + 浴室数量：绝大多数为1个。
  + 卧室数量：主要为1个，少量为2个。
  + 面积：平均750平方英尺，相对较小。
* **簇1 (12%)：大户型公寓**
  + 浴室数量：主要为2个，也有更多的情况。
  + 卧室数量：主要为3个，也有更多或更少的情况。
  + 面积：平均1463平方英尺，相对较大。
* **簇2 (32%)：中等户型公寓**
  + 浴室数量：几乎全部为2个。
  + 卧室数量：几乎全部为2个。
  + 面积：平均1096平方英尺，中等大小。

**4. 可视化分析**

我们使用WEKA的“Visualize Cluster Assignments”功能，对聚类结果进行了可视化分析，选择了以下几组属性组合：

1. **bedrooms (X轴) vs. bathrooms (Y轴)**：观察卧室数量与浴室数量的关系。结果显示，总体上呈现正相关趋势，且不同簇大致呈现出沿着对角线分布的形状，分别对应1卧室1浴室、2卧室2浴室、以及3卧室2浴室或更多的组合。
2. **square\_feet (X轴) vs. bedrooms (Y轴)**：观察面积与卧室数量的关系。结果显示，总体上呈现正相关趋势，不同簇在图中也形成了较为明显的区域划分。

EM算法成功地将公寓出租信息数据集划分为三个具有明显特征的簇，分别对应小户型、中等户型和大户型公寓。聚类结果与我们的常识和预期基本一致，验证了聚类分析的有效性。

需要注意的是，EM算法只进行了1次迭代就收敛了，这可能是因为初始化的聚类中心已经比较接近最优解，或者数据本身就比较容易分离。在未来的工作中，可以尝试增加迭代次数、使用不同的随机种子、或引入其他特征，进一步探索和优化聚类结果。

**五、规则挖掘**

本部分旨在利用关联规则挖掘技术，进一步探索公寓出租信息数据集中各属性间的潜在关系。这里采用WEKA数据挖掘平台提供的Apriori算法，对经过预处理的数据进行分析。Apriori算法是关联规则挖掘的经典方法，能够高效地发现满足预设支持度和置信度阈值的频繁项集及关联规则。此部分摘除了房屋的地理位置（城市+州）属性，单独讨论房屋本身属性间是否存在规则关联。

**1. 实验设置**

* **工具：** WEKA (Version 3.8.6)
* **算法：** Apriori
* **数据集：** 经过预处理的公寓出租信息数据集（apartment+for+rent+classified+100k-weka.filters...）
* **属性：**
  + bathrooms (浴室数量，已离散化)
  + bedrooms (卧室数量，已离散化)
  + pets\_allowed (是否允许宠物，已离散化)
  + price (价格，已离散化)
  + square\_feet (面积，已离散化)
* **Apriori算法参数：**
  + -N 5: 生成的规则数量上限（本例中设置为5）。
  + -T 0: 使用置信度（Confidence）作为规则评估指标。
  + -C 0.9: 最小置信度阈值（0.9）。
  + -D 0.05: 置信度增量（用于迭代搜索最佳置信度）。
  + -U 1.0: 支持度上限（1.0，即不设上限）。
  + -M 0.1: 最小支持度阈值（0.1，但实际最小支持度为0.35，见下文）。
  + -S -1.0: 重要性度量（未使用）。
  + -c -1: 类索引（未使用，因为本实验不是分类任务）。

**2. 实验结果**

WEKA Apriori算法的运行结果如下：

=== Run information ===

Scheme: weka.associations.Apriori -N 5 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: apartment+for+rent+classified+100k-weka.filters.unsupervised.attribute.Remove-R1,7-8-weka.filters.unsupervised.attribute.Discretize-F-B5-M-1.0-R4,5-precision6-weka.filters.unsupervised.attribute.Discretize-F-B5-M-1.0-R5-precision6-weka.filters.unsupervised.attribute.Discretize-B9-M-1.0-R1-precision0-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-R2-precision0

Instances: 99972

Attributes: 5

bathrooms

bedrooms

pets\_allowed

price

square\_feet

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.35 (34990 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 13

Generated sets of large itemsets:

Size of set of large itemsets L(1): 5

Size of set of large itemsets L(2): 5

Size of set of large itemsets L(3): 1

Best rules found:

1. bedrooms='(1-2]' pets\_allowed=Cats,Dogs 37691 ==> bathrooms='(-inf-2]' 37456 <conf:(0.99)> lift:(1.66) lev:(0.15) [14904] conv:(64.15)

2. bedrooms='(1-2]' 40663 ==> bathrooms='(-inf-2]' 40407 <conf:(0.99)> lift:(1.66) lev:(0.16) [16077] conv:(63.55)

3. bathrooms='(2-3]' 38853 ==> pets\_allowed=Cats,Dogs 37209 <conf:(0.96)> lift:(1.03) lev:(0.01) [1097] conv:(1.67)

4. bedrooms='(2-3]' 46206 ==> pets\_allowed=Cats,Dogs 43495 <conf:(0.94)> lift:(1.01) lev:(0.01) [548] conv:(1.2)

5. bathrooms='(-inf-2]' bedrooms='(1-2]' 40407 ==> pets\_allowed=Cats,Dogs 37456 <conf:(0.93)> lift:(1) lev:(-0) [-100] conv:(0.97)

**3. 结果分析与讨论**

Apriori算法在给定的参数设置下，找到了5条满足最小支持度（0.35）和最小置信度（0.9）的关联规则。下面对这些规则进行逐条分析：

1. **bedrooms='(1-2]' pets\_allowed=Cats,Dogs 37691 ==> bathrooms='(-inf-2]' 37456 conf:(0.99) lift:(1.66)**
   * 这条规则表明，在拥有1-2间卧室且允许猫和狗的公寓中，有99%的概率浴室数量不超过2间。
   * 提升度（Lift）为1.66，说明这条规则的出现概率远高于两个条件独立出现的情况，具有较强的关联性。
2. **bedrooms='(1-2]' 40663 ==> bathrooms='(-inf-2]' 40407 conf:(0.99) lift:(1.66)**
   * 这条规则表明，在拥有1-2间卧室的公寓中，有99%的概率浴室数量不超过2间。
   * 提升度为1.66, 与上一条类似。
3. **bathrooms='(2-3]' 38853 ==> pets\_allowed=Cats,Dogs 37209 conf:(0.96) lift:(1.03)**
   * 这条规则表明，在拥有2-3间浴室的公寓中，有96%的概率允许猫和狗。
   * 提升度略大于1，关联性较弱，但仍然表明两者存在正相关。
4. **bedrooms='(2-3]' 46206 ==> pets\_allowed=Cats,Dogs 43495 conf:(0.94) lift:(1.01)**
   * 这条规则表明，在拥有2-3间卧室的公寓中，有94%的概率允许猫和狗。
   * 提升度接近于1，关联性较弱。
5. **bathrooms='(-inf-2]' bedrooms='(1-2]' 40407 ==> pets\_allowed=Cats,Dogs 37456 conf:(0.93) lift:(1)**
   * 这条规则置信度较高, 但提升度为1, 表明这两个条件几乎是独立发生的。

**总结：**

从挖掘出的关联规则可以看出：

卧室数量和浴室数量之间存在较强的关联性，尤其是1-2间卧室的公寓，浴室数量通常不超过2间。而允许饲养宠物（猫和狗）与公寓的卧室、浴室数量之间存在一定的关联，但关联强度不如卧室与浴室之间的关联。另外，规则5虽然置信度较高，但提升度为1，说明这两个条件（浴室数量少于等于2，卧室数量在1-2间）和允许猫狗这个结果是独立同分布的。